Measuring Word Learning Performance in Computational Models and Infants

Christina Bergmann*†, Lou Boves*, Louis ten Bosch*
*Centre for Language & Speech Technology, Radboud University, Nijmegen, The Netherlands
†International Max Planck Research School for Language Sciences, Radboud University, Nijmegen, The Netherlands

***Abstract*—In the present paper we investigate the effect of categorising raw behavioural data or computational model responses. In addition, the effect of averaging over stimuli from potentially different populations is assessed. To this end, we replicate studies on word learning and generalisation abilities using the ACORNS models. Our results show that discrete categories may obscure interesting phenomena in the continuous responses. For example, the finding that learning in the model saturates very early at a uniform high recognition accuracy only holds for categorical representations. Additionally, a large difference in the accuracy for individual words is obscured by averaging over all stimuli. Because different words behaved differently for different speakers, we could not identify a phonetic basis for the differences. Implications and new predictions for infant behaviour are discussed.**

I. Introduction

In language acquisition research, be it child experiments or computational simulations, one must decide how to measure and report behaviours. Most research relies on accuracy measures, averaged over groups of stimuli or participants. However, this averaging might both obscure potentially relevant effects and give rise to false assumptions. [1] showed that an overall average, implying the assumption that participants and stimuli in (psycho)linguistic experiments are drawn from essentially homogeneous populations, may lead to potentially misleading conclusions about the processes that yielded the observation data. Participants and/or stimuli may actually come from different populations that show different, yet systematic behaviours, which may be masked by only investigating mean values [2].

For example, [3] found that while infants can *discriminate* the minimally different sound pattern [bɪn] and [dɪn], they seemed unable to distinguish those two syllables in a word-learning task. However, upon closer inspection of the data, Fikkert found that her participants succeeded in half of the learning task: they noticed the change from the familiar '<u>b</u>in' to the new word '<u>d</u>in', but the reverse case appeared to go unnoticed. By reporting only gross average results, the success for the '<u>b</u>in' to '<u>d</u>in' change is not detectable.

[4] reported interesting insights into the errors children make during pronoun comprehension. A separate analysis of test items by (1) verb frequency and (2) gender of the pronoun revealed diverging performances across both factors. In previous studies, verb frequency and pronoun gender have not been analysed separately and were thought to have no influence on children's performance. This assumption was, however, solely based on theoretical accounts and only supported by group averages. A difference in performance based on verb frequency and pronoun gender is a challenge for all current models of pronoun comprehension and acquisition.

The assumption that participants and items are drawn from a single homogeneous population is not the only problem. In language acquisition experiments the behaviours of the participants are often quantified in two categories: correct (for example if the infant looks in the predicted direction) or wrong (if the infant looks in a different direction). Although such a binary classification is intuitively appealing, it is to a large extent arbitrary. Collapsing 'not looking in any specific direction' and fixating on the 'wrong' picture in a preferential looking task may or may not be warranted. Also, infants seldom fixate exclusively on a single part of the screen; rather, they focus on a picture for 'most of the time'. Here, too, a binary choice may ignore relevant data, which might indicate the confidence of the participant in reacting to a specific stimulus. The time course of the fixations also contains important information, which cannot be accounted for in a 'correct/incorrect' classification of a response.

In this paper we address the interaction between the representation of the behavioural data used in statistical analyses and the details of the analyses. To this end, we first investigate whether binary scoring of responses suggests different conclusions than a representation that retains information about the confidence of a participant. In this part of the research we still use mean values computed over all test items. Second, we investigate whether different ways of quantifying the data may have an impact on the assumption that all stimuli in an experiment come from a homogeneous population by comparing item-based analyses to assessments based on averages.

Rather than trying to re-analyse behavioural data from infant experiments, we investigate the issue by means of computational simulations. The output of computational models of language acquisition (see [5] for a summary of recent models) can be (semi-)continuous functions, like the proportion of time an infant fixates on parts of a picture in a looking-while-listening experiment. For example, some of the models investigated in [5] return activations, but to report on the models' performance in terms of precision and recall, discretising thresholds are applied.

In this paper, we examine one of the models developed in the ACORNS (ACquisition Of Recognition and communica-

978-1-61284-990-4/11/$26.00 ©2011 IEEE

tioN Skills) project[1]. The model uses real speech and simulated visual representations of a scene as input and assumes only general cognitive (learning) abilities [6]. In the ACORNS project it was shown that it is possible to learn associations between speech utterances and visual representations of objects without the need to first segment the speech signal into phone-like units. Furthermore, experiments suggested that few exposures to utterances containing a given word, paired with a label to represent cross-modal (visual) input, were sufficient to obtain a very high overall recognition accuracies.

The ACORNS model has been used to simulate learning of up to 25 content words. For each test stimulus the model returns an activation value for all the words that are being learned. Previous experiments with the model almost invariably categorised the 'raw' output of the model into two response categories: *correct* or *incorrect*. In this paper we replicate model simulations [6] and investigate to what extent the results and the interpretation of the simulations are influenced if the 'raw' (continuous, real-valued) activations are analysed directly, omitting the binary classification. In addition, we investigate whether there are differences between speakers from which the model learns and between target words (*keywords*) to examine whether a more detailed analysis confirms the assumption of a homogeneous performance.

To assess the impact of the representation of the model's output and of the way that the 'raw' model output is used in subsequent statistical processing, we replicated computational simulations that address a specific question in language acquisition: does a larger degree of variation during learning aid generalisation? In particular, does learning from one or from multiple speakers affect recognition of unknown talkers [7], [8]? Previous simulations with the ACORNS model in an experiment in which nine keywords were learned seemed to corroborate the conclusion of behavioural experiments that learning from multiple speakers facilitates recognition of new speakers more than learning from a single speaker [6], [9].

II. Experiments

A. Non-Negative Matrix Factorisation

Learning in the ACORNS model is based on Non-negative Matrix Factorisation (NMF) [10]. NMF simulates learning by finding a decomposition of an $n \times m$ dimensional input matrix $\mathbf{V}$, consisting of m utterances, each encoded as a vector $\vec{v}$ of dimension n (representing the acoustic features $\vec{v}_a$ and conceptual keyword labels $\vec{v}_k$ of an utterance). NMF decomposes this matrix into the product of two smaller matrices $\mathbf{W} \cdot \mathbf{H} \approx \mathbf{V}$ by minimising the Kullback-Leibler divergence between the input $\mathbf{V}$ and the product of $\mathbf{W}$ and $\mathbf{H}$. The dimension of $\mathbf{W}$ is $n \times r$, and the dimension of $\mathbf{H}$ is $r \times m$. The constant r is chosen such that $(m + n)r \ll m \times n$, i.e., information is compressed. In the present experiments, r equals 70, which means the model has ample space to accommodate all learned keywords internally.

[1] http://www.acorns-project.org

The matrix $\mathbf{W}$ has the same structure as the input in $\vec{v}$, namely an acoustic and a conceptual keyword encoding part. Hence, each column vector in $\mathbf{W}$ can be considered as representing an association between acoustic and semantic information of keywords. Thus, $\mathbf{W}$ contains the internal representations that emerge during learning. $\mathbf{H}$ contains information about activation of columns in $\mathbf{W}$ during training. We used an incremental version of NMF [10], which only needs to memorise the most recent utterances in addition to the internal representations in the matrix $\mathbf{W}$.

During testing, a new utterance is given to the model in the same acoustic encoding $\vec{v}_a$ as in the training [11], but without providing the corresponding keyword part $\vec{v}_k$. The missing keyword information has to be reconstructed by the learner based on the stored internal representations. This can be done by approximating $\vec{v}_k \approx \mathbf{W}_k \cdot \hat{h}$ (via minimising the Kullback-Leibler divergence), where $\hat{h}$ is estimated using only the acoustic information in the learned representations in $\mathbf{W}$. The activation values of the keyword labels in the reconstructed vector $\hat{h}$ take real values, unlike the binary keyword labels $\vec{v}_k$ presented to the learner in training. During testing the incremental learning is switched off, so that the processing of test utterances does not affect the internal representations. Thus, the same test stimuli can be used repeatedly to track the progress of the learning process.

B. Response Scoring

To take advantage of the fact that each test utterance yields a vector of real-valued activations instead of a binary 'correct/wrong' value, we assessed the model's recognition performance in terms of activations for all nine keywords. The activation can be interpreted as the confidence of the learner's associations of acoustic stimuli with its internal word representations.

The performance of the model was assessed by generating a confusion matrix. In each experiment, two such matrices were generated. The first matrix accumulated the the number of times that a given stimulus was recognised as one of the nine keywords. Thus, the diagonal of this matrix contains the number of times a keyword was recognised correctly. Dividing the counts by the number of test utterances yields proportional accuracy. In the remainder of the paper we will indicate the results obtained with this matrix as *crisp accuracy*. In the second matrix we accumulated the normalised activations of all keywords given a specific test sentence. The values on the main diagonal of this matrix will be referred to as *fuzzy accuracy*. In the fuzzy measurement, values can only be 1 (or 100%) if all test tokens of a given keyword receive activations from only one column in the learner-internal $\mathbf{W}$ matrix.

C. Training and Testing

As in previous experiments (e.g., [6]), we investigate whether learning from multiple speakers aids generalisation to new speakers more than learning from a single speaker. For this purpose we train the learner in two conditions. In the first, the model learns from four speakers in a row. In

the second condition, the model learns from all four speakers intermixed. These two conditions are termed *speaker-blocked* and *speaker-mixed*. The model had to learn nine different keywords, which were always embedded in short but varying carrier sentences. These sentences were recorded in a virtually noise-free environment by four native speakers of English, two male and two female. Each keyword occurred 60 times spoken by one of the four speakers (for a total of 540 utterances per speaker, summing to 2160 utterances in total). These sentences were identical across conditions and merely presented to the learner in different sequences.

In the speaker-blocked condition the model was trained with one speaker at a time. Thus, the model first experienced no speaker variation and only learned from utterances spoken by a single speaker. The training utterances were ordered in such a way that each block of nine utterances contained all nine words to be learned. The carrier sentences in which the words were embedded were randomised. When all 540 utterances of the first speaker were processed, a second speaker was used for training, thereby increasing introducing additional variation in the model's input. This held then, too, for the onset of the third and fourth speaker. In the speaker-mixed condition, learning stimuli were randomised across speakers so that the model could learn from all four speakers simultaneously.

To test the model's recognition performance, a fixed held-out test set was used containing 20 utterances per keyword and speaker. This test set was identical for both training conditions. During testing, the model was frozen in its current state and thus could not learn from being exposed to the test utterances or to new speakers. Testing was conducted independently for each speaker, so that the difference in performance for yet unknown speakers could be assessed in the speaker-blocked condition. Thus, we took full advantage of this blocked presentation of speakers during training and could assess the model's generalisation abilities to unknown speakers when it has observed one to three speakers.

To get insight into the model's behaviour as training proceeds, we tested the model at regular time intervals during learning. In the beginning of the learning stage, the model was tested at every tenth utterance for 90 utterances from the point of a speaker change onwards. For the remainder of the training, we assessed the model's responses after a new block of 90 training utterances had been observed.

D. Research Questions

The expectation regarding performance is that fuzzy accuracy is lower than the crisp, count-based measurement used in previous studies. This is due to the fact that the crisp accuracy assumes that the representation with the highest activation is uniquely selected, effectively reducing the activation of competing representations to zero. Furthermore, and more important here, the shape of the fuzzy learning curve will not only be determined by the number of test items recognised correctly, but also by the strength of activation and thus the amount of competition during recognition.

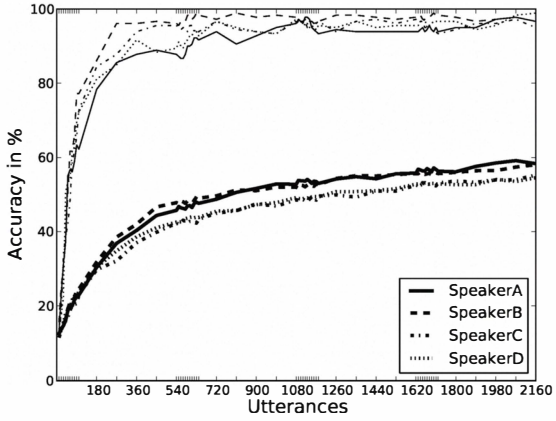

Fig. 1: Crisp and fuzzy accuracy results for all four speakers in the speaker-mixed condition. The upper lines denote crisp accuracy, the lower the fuzzy measurement. Speaker identities are consistently denoted with the given line styles in thick (fuzzy accuracy) and thin (crisp accuracy) lines.

We expect that a more detailed analysis uncovers possible differences between speakers, specifically in the speaker-blocked condition. Regarding the group of keywords, the implicit assumption of previous studies [6] was that they stem from a homogeneous group and to not elicit systematically different responses.

Previous studies showed that both training schemes, namely speaker-mixed versus speaker-blocked training, reach ceiling accuracy at or near 100% in the crisp measurement quickly [6].

III. Results

A. Crisp Accuracy

1) Speaker-Mixed: Conventional, crisp accuracy indicates very high performance across training for the speaker-mixed case (top lines of Fig. 1). Each speaker seems to be recognised equally well and the learning curves are comparably steep. Accuracies above 80% are reached within 180 utterances, after 540 utterances all speakers reach accuracies above 90%.

2) Speaker-Blocked: For the speaker-blocked condition, depicted in Fig. 2, it appears that the speaker currently trained quickly reaches very high accuracy performance in the range of $98-100\%$. Speakers not yet trained seem to profit to some extent from training with one speaker and even more so from switching speakers. However, the size of the effect seems to differ for each speaker. Another notable result in the speaker-blocked condition is the decrease in accuracy after a speaker is no longer used for training. Most notably, the first speaker drops to an accuracy level of about 70% at the end of training.

To facilitate comparability with previous studies, Fig. 2 also contains the average recognition accuracy across all speakers (square-marked solid line), which was used to report the model's performance in [6]. This measure shows steadily increasing performance with each speaker change. Furthermore, the overall average can be used to compare the model's

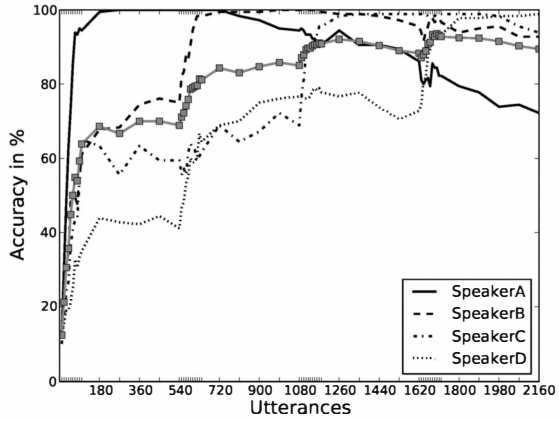

Fig. 2: Crisp accuracy performance in the speaker-blocked condition for each of the four speakers. Speaker changes occur every 540 utterances, in alphabetical speaker order. The average performance over all speakers is denoted in the square-marked grey line plot.

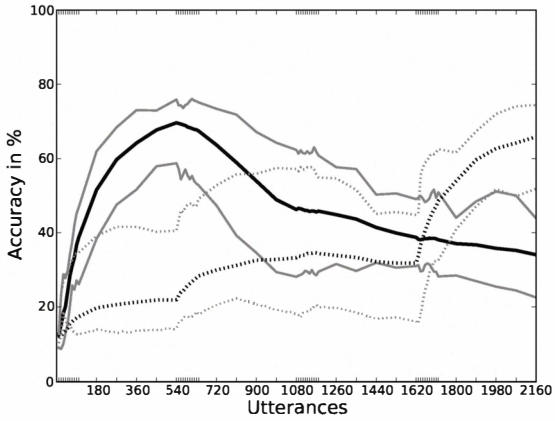

Fig. 3: Performance of the first and last speaker trained in the speaker-blocked condition based on the fuzzy accuracy. Additionally, the highest and lowest performing keyword for each speaker is depicted to illustrate the range of keyword performances contributing to the overall learning curves for each speaker.

performance across conditions. Since the difference between speakers in the speaker-mixed condition is small, the overall average is at the same level as speaker accuracy. A comparison of both averages over speakers shows that the model performs better in the speaker-mixed condition than in the speaker-blocked condition throughout training.

B. Fuzzy Accuracy

1) Speaker-Mixed: When comparing the two different scoring systems described in Sec. II-B, the expectation of overall lowered performance in terms of relative activation is confirmed, as shown in Fig. 1. This plot depicts the performance in the speaker-mixed condition, both according to crisp and fuzzy accuracy measurements. It can be seen that all speakers perform on a similar level within one measurement. However, while the crisp measurement indicates a ceiling effect after only 180 training utterances have been observed, the fuzzy assessment shows that activations for the correct keyword label continue to increase throughout learning. This is evident in the ongoing upward trend, starting from 30% at utterance 180 and roughly doubling activation at the end of training for all speakers.

2) Speaker-Blocked: The performance for two of the speakers in the speaker-blocked condition is shown in Fig. 3; performance for the remaining two speakers was comparable and is omitted to ensure clarity of the figure. Comparing the first and last speaker in training, Speaker A and Speaker D in Fig. 2 and in Fig. 3, the finding that there is no ceiling effect is confirmed across conditions.

Additionally, the beneficial effect of training with more than one speaker in the speaker-blocked condition in the crisp measurement almost vanishes in the fuzzy assessment. Thus, the model seems to increase the amount of correct responses, albeit with only slight changes in the activation pattern. This seemingly marginal shift has a severe impact on crisp accuracy performance (cf. the increase from 40% to 70% correct for Speaker D between utterances 540 and 810 in Fig. 2 compared to an increase of 10% in the same time frame for the same speaker in Fig. 3).

C. Keyword Accuracy

When analysing the performance for single speakers and keywords, differences emerge that were not predicted from the overall accuracy scores reported in the previous sections. As an example of the variation between keywords, Fig. 3 depicts the highest and lowest performing keyword for Speaker A and Speaker D, the first and last speaker trained in the speaker-blocked condition. The figure shows that different keywords can be affected in various ways by the blocked training. First, the increase in activation strength during training with the same speaker varies greatly across speakers and keywords. Second, in the fuzzy accuracy condition the decrease when a speaker is no longer observed during training and the model learns from new speakers, seems to be very steep and sudden for Speaker A's lowest performing keyword, but more gradual for the best keyword. Third, the beneficial effect of variation on Speaker D seems more pronounced for the high performing keyword, whereas the low performing keyword does not seem to be affected until that speaker is used for training. To summarise, not all keywords reach the same level of either gross accuracy or relative activation strength, showing a difference of internal representations and a non-uniform effect of training.

These findings hold not only across speaker-blocked and mixed conditions, but also for both types of model assessment. Using crisp accuracy, as depicted in Fig. 4, the range of keyword performances is even wider than in the fuzzy assessment

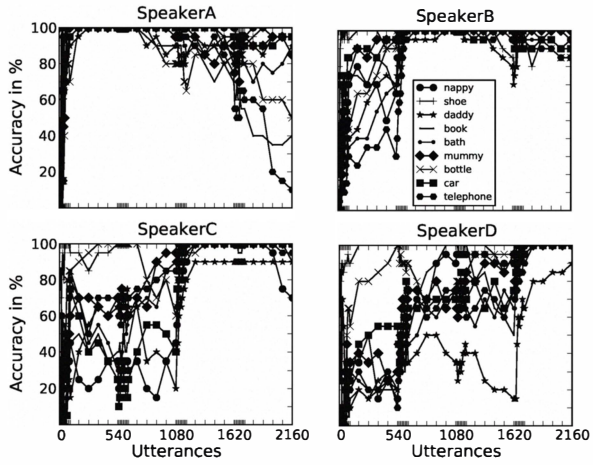

Fig. 4: Overview over the crisp recognition performance for each keyword per speaker in the speaker-blocked condition. Speaker were trained in alphabetical order with changes occurring every 540 utterances. Each line denotes the crisp accuracy for one of the keywords spoken by a single speaker.

depicted in Fig. 3, when a speaker is not currently trained in the speaker-blocked condition. When a speaker is currently not used for training, the range of keyword performances extends from chance level to perfect accuracy at 100%. Again, the identity of those keywords varies across speakers and conditions. The overview given in Fig. 4 illustrates the high variability according to speakers and between keywords, which is only diminished when the respective speaker is currently used to train the model.

IV. Discussion

The most salient results are that (1) two types of training schemes (speaker-blocked and speaker-mixed) led to high accuracies, albeit with different final performance. Gross averages showed that the speaker-mixed condition led to a higher final accuracy than the speaker-blocked condition. This was on one hand due to decreasing accuracies for speakers trained early on and on the other hand caused by low accuracies for speakers not yet trained in the speaker-blocked condition. These speaker-dependent differences only showed up in a more detailed examination.

From the fuzzy accuracy measurement, which is closer to the model's actual output and requires almost no post-processing, it became apparent that the high crisp accuracy values do not necessarily coincide with saturated learning. Even though the model performs at a high crisp accuracy level, more training led to increased fuzzy values.

A. Crisp Accuracy

The learner could profit from variation in the input in two ways. In the speaker-blocked condition yet unknown speakers performed above chance level and increased in recognition accuracy when more than one speaker has been observed by the model. In addition, the performance for all speakers in the speaker-mixed condition was above the average performance in the speaker-blocked condition. This means that the model could learn from all four speakers at the same time and was able to build representations that led to high recognition accuracies for all trained speakers at the same time in the speaker-mixed condition. The speaker-blocked condition could not reach an overall comparable level of accuracy. However, the speaker currently being trained reached almost perfect accuracy in the speaker-blocked condition, thus outperforming the speaker-mixed condition when considering single-speaker performance. This shows that only observing one speaker at a time leads to a better adaptation to that given speaker, whereas more variant training allows for high recognition performance for a greater range of stimuli.

B. Fuzzy Accuracy

Taking into account the model's real-valued output and inspecting what was termed *fuzzy accuracy* in the present paper, it became apparent that learning was not saturated even though the model achieved very high crisp recognition accuracies. This is evident from the continuing increase in fuzzy accuracy, denoting a sharpened activation pattern, long after ceiling performance is reached in the crisp accuracy assessment (e.g., Fig. 1). This finding points towards the model's ability to recognise the correct keyword even in uncertain circumstances. This is most pronounced for the improvements of yet unseen speakers in the speaker-blocked condition. While crisp accuracy shows a sharp rise for Speaker D, the last speaker trained, when the second speaker comes in (Fig. 2), the fuzzy measurement showed that activations remained on a comparatively low level (Fig. 3). This shows that confidence and accuracy are genuinely different measurements of performance that can lead to diverging impressions.

C. Keyword Accuracy

When further investigating possible effects on the level of keywords, a high variation in performance was uncovered. This is most evident in the speaker-blocked condition, as depicted in Fig. 4 for the crisp accuracy measurement. Some words only reached a recognition accuracy on chance level, while others were recognised in almost all tests throughout training. Averaging performance over items, as is the case in linguistic studies [2], obscures this high variability. However, we did not discover distinct groups of keywords that seemed to fall into two or more categories. Rather, performance was spread out over a large range. There was also only little systematicity regarding which keywords performed at a high or low level. Thus, we did discover that single keywords do not necessarily perform at the same level, but we could not discover inherent properties of the keywords that lead to fundamentally distinct recognition performance.

D. Word Learning in Models and Infants

To relate this finding to (psycho)linguistic studies, a binary assessment of response data (e.g., fixations, answers) can

yield an incomplete picture. By considering the graded nature of responses, such as reaction times or fixation duration, new insights might be gained. The continuing increase in activation found in the model, even after a very high crisp accuracy level had been reached, might correspond to decreased processing cost with increased exposure to a given keyword in infants. When analysing known words according to their frequency, a difference in dynamic measurements such as fixation data or reaction times of overall highly accurate responses should emerge. Such frequency effects have been observed in adults [12] and might be at work in children as well, even though they are only beginning to acquire their native language.

Another finding regards the potentially beneficial effect of variation on a learner's generalisation abilities [7], [8], [13]. Both the speaker-mixed and the speaker-blocked condition let the model learn from the same set of utterances spoken by four different speakers, providing the learner eventually with the same amount of variation, merely structured differently. Our results, both averaged and on the speaker- and keyword-level, show that the type of short-range variation influences general recognition accuracy. The blocked training scheme led to overall worse performance than the mixed presentation of speakers, with decreasing recognition accuracies for the first speakers after they were no longer trained. This finding points towards two opposing tendencies when a learner has to build representations that allow for reliable word recognition: On one hand, adaptation to a single speaker favours more specific representations, whereas on the other hand broader, less specific representations allow for increased generalisability. These two trends come to bear in different ways in the two training schemes presented here.

Based on the results presented above, we would expect that in (psycho)linguistic studies infants' ability to generalise is modulated by the order of presentation of the learning stimuli. Depending on whether speakers change at every item or are blocked, infants should show a difference in recognition behaviour. Only the truly mixed case has so far been tested systematically (c.f. [7], [8]) and this evidence supports the claim that variation drives generalisation abilities (without specific requirements regarding this variation), whereas a more blocked variation of speakers seems to be more natural in the infants' input.

The difference between the effects of blocked and intermixed presentation on performance can affect infant studies at a very different point as well. Consider paradigms relying on habituation or familiarisation, such as the Headturn Preference Procedure [8], [14], where possibly variant stimuli can be presented in an intermixed or blocked order in the first phase. These two possibilities might, according to our results, affect infant's behaviour in the test phase. This, however, has to be assessed systematically to first verify that the effect is present in children and to second quantify where it applies and has measurable consequences.

However, all predictions can only be made with caution, as some of our findings might be due to the way learning is simulated in this specific model and the way acoustic and visual information is presented to the learner, as described in Sec. II-A. To be able to fully generalise the present findings to infant behaviour, future investigations into the model's inner workings and the analogies in infants are necessary.

E. Conclusion

The present study compared two ways of assessing performance of a computational word learner and showed that considering only categorised output might lead to wrong conclusions. In this specific case, we found that learning was not saturated in terms of activations of the correct response. At the same time, the overall, crisp performance was at ceiling. In addition, averaging accuracy over keywords could be justified in the current study, because we could not uncover distinct keyword sub-populations. However, averaging over speakers obscured effects in one condition, namely when presenting speakers in a blocked fashion. There, both the decreasing accuracy after training and the only partial gain from variation for yet unknown speakers was not visible in the overall data. These findings underline the need to assess underlying data and justify when and when not to report only average results.

ACKNOWLEDGEMENTS

The research of Christina Bergmann is supported by grant number 360-70-350 from the Dutch Science Organisation NWO.

REFERENCES

[1] H.H. Clark, "The language-as-fixed-effect fallacy: A critique of language statistics in psychological research", *Journal of Verbal Learning and Verbal Behavior*, 12: 335–359, 1973.

[2] K. Forster and M. Masson, "Introduction: Emerging data analysis [editorial]", *Journal of Memory and Language (Special Issue: Emerging Data Analysis)*, 59(4): 387–388, 2008.

[3] P. Fikkert, "Developing representations and the emergence of phonology: Evidence from perception and production", *Laboratory Phonology 10*, 227–260, 2010.

[4] D. Matthews, E. Lieven, A. Theakston, and M. Tomasello, "Pronoun co-referencing errors: Challenges for generativist and usage-based accounts", *Cognitive Linguistics*, 20(3): 599–626, 2009.

[5] B. Macwhinney, "Computational models of child language learning: an introduction [editorial]", *Journal of Child Language*, 37: 477–485, 2010.

[6] L. ten Bosch, H. Van hamme, and L. Boves, "Unsupervised detection of words – questioning the relevance of segmentation", *ISCA ITRW, Speech Analysis and Processing for Knowledge Discovery*, paper 046, 2008.

[7] R.S. Newman, "The level of detail in infants' word learning", *Current directions in Psychological Science*, 17(3): 229–232, 2008.

[8] D.M. Houston and P.W. Jusczyk, "The role of talker-specific information in word segmentation by infants", *Journal of Experimental Psychology: Human Perception and Performance*, 26(5): 1570–1582, 2000.

[9] C. Bergmann, M. Gubian, and L. Boves. "Modelling the effect of speaker familiarity and noise on infant word recognition", *Proceedings of Interspeech 2010*, 2910–2913, 2010.

[10] J. Driesen, L. ten Bosch, and H. Van hamme, "Adaptive Non-negative Matrix Factorization in a Computational Model of Language Acquisition", *Proc. Interspeech 2009*, 1731–1734, 2009.

[11] H. Van hamme, "HAC-models: a novel approach to continuous speech recognition", *Proc. Interspeech 2008*, 2554–2557, 2008.

[12] J. Zevin, "Word Recognition", In: Larry R. Squire, *Encyclopedia of Neuroscience*, Academic Press, Oxford, 517–522, 2009.

[13] R.L. Gómez, "Variability and detection of invariant structure", *Psychological Science*, 13(5): 431–436, 2002.

[14] P.W. Jusczyk and R.N. Aslin, "Infants Detection of the Sound Patterns of Words in Fluent Speech", *Cognitive Psychology*, 29(1): 1–23, 1995.